

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-051890

(43)Date of publication of application : 23.02.2001

(51)Int.Cl.

G06F 12/00

G06F 13/00

(21)Application number : 11-226494

(71)Applicant : TOSHIBA CORP

(22)Date of filing : 10.08.1999

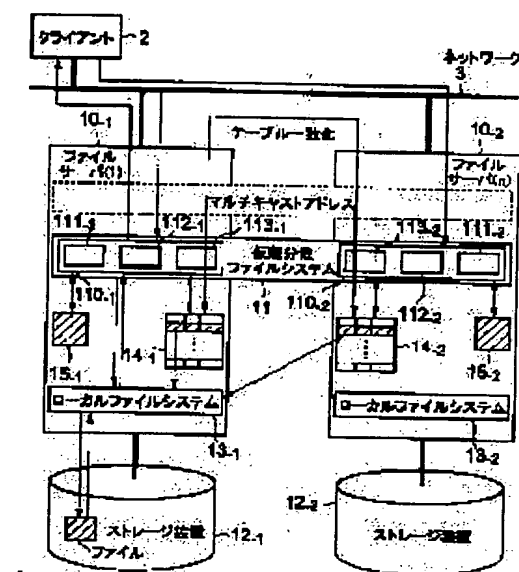
(72)Inventor : UCHIBORI IKUO
TAKAKUWA MASAYUKI

(54) VIRTUAL DECENTRALIZED FILE SERVER SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To make a client not pay attention to the number of file servers decentralized in a network and the connection states of storage devices.

SOLUTION: This virtual decentralized file server system 1 is equipped with servers 10-1 and 10-2 decentralized in the network 3 and a virtual decentralized file system 11 is decentralized and mounted on each of the servers. Modules 110-1 and 110-2 on the servers 10-1 and 10-2 which constitute this system 11 when receiving a file operation request multicast from a client 2 judge whether or not their servers are optimum servers capable of handling the request according to server information holding parts 15-1 and 15-2 holding mapping tables 14-1 and 14-2 between the virtual decentralized file system 11 and all local file systems 13-1 and 13-2 or server information on all the servers, and makes a local file system of a corresponding server perform requested file operation according to the judgement result.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-51890

(P2001-51890A)

(43) 公開日 平成13年2月23日 (2001.2.23)

(51) Int.Cl. ⁷	識別記号	F I	テーマコード* (参考)
G 0 6 F 12/00	5 4 5	G 0 6 F 12/00	5 4 5 A 5 B 0 8 2
13/00	3 5 1	13/00	3 5 1 E 5 B 0 8 9

審査請求 未請求 請求項の数 7 O L (全 14 頁)

(21) 出願番号 特願平11-226494
(22) 出願日 平成11年8月10日 (1999.8.10)

(71) 出願人 000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72) 発明者 内堀 郁夫
東京都府中市東芝町1番地 株式会社東芝
府中工場内
(72) 発明者 高桑 正幸
東京都府中市東芝町1番地 株式会社東芝
府中工場内
(74) 代理人 100058479
弁理士 鈴江 武彦 (外6名)

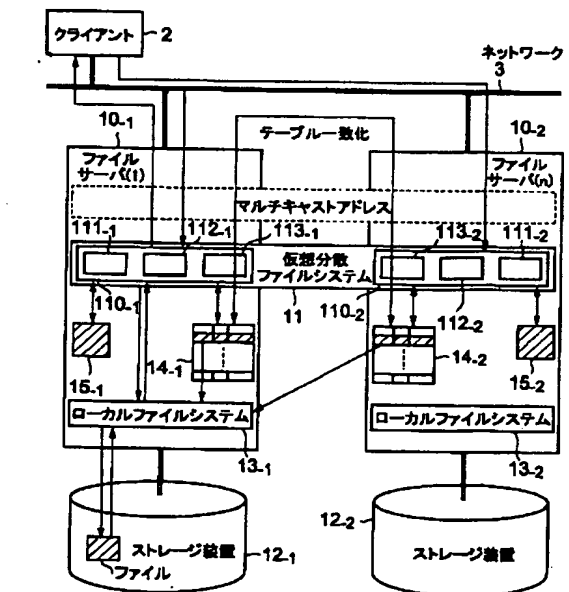
最終頁に続く

(54) 【発明の名称】 仮想分散ファイルサーバシステム

(57) 【要約】

【課題】 ネットワーク上に分散した複数のファイルサーバの台数やストレージ装置の接続状態をクライアントに意識させないで済むようにする。

【解決手段】 ネットワーク 3 上に分散したサーバ 10-1, 10-2 を備え、各サーバには、仮想分散ファイルシステム 11 が分散して実装されている。このシステム 11 を構成する、サーバ 10-1, 10-2 上のモジュール 110-1, 110-2 は、クライアント 2 からマルチキャストされたファイル操作要求を受け取ると、仮想分散ファイルシステム 11 と全ローカルファイルシステム 13-1, 13-2 とのマッピングテーブル 14-1, 14-2 または全サーバのサーバ情報を保持するサーバ情報保持部 15-1, 15-2 をもとに、自サーバが上記要求を処理可能な最適なサーバであるか否かを判断し、その判断結果に基づいて要求されたファイル操作を対応するサーバのローカルファイルシステムにより行わせる。



1
仮想分散ファイル
サーバシステム
14-1, 14-2...マッピングテーブル
15-1, 15-2...サーバ情報保持部
110-1, 110-2...仮想分散ファイルモジュール
111-1, 111-2...仮想分散ファイルインタフェース
112-1, 112-2...ローカルファイルインタフェース
113-1, 113-2...通信モジュール

【特許請求の範囲】

【請求項 1】 マルチキャスト可能なネットワーク上に分散した複数のファイルサーバを備えた仮想分散ファイルサーバシステムであって、

前記各ファイルサーバに分散して実装され、全ファイルサーバのファイルを統合的に管理する、実際のストレージ構成には非依存の仮想分散ファイルシステムと、

前記各ファイルサーバにそれぞれ独立して実装され、各サーバに固有のストレージ構成を管理するローカルファイルシステムと、

前記各ファイルサーバにそれぞれ設けられ、前記仮想分散ファイルシステムで統合的に管理される各ファイルについて、当該仮想分散ファイルサーバシステムとそのファイルを実際に管理する前記ローカルファイルシステムとの間のマッピングの情報を保持するマッピングテーブルとを具備し、

前記仮想分散ファイルシステムは、前記各ファイルサーバにそれぞれ設けられた管理モジュールから構成され、前記各管理モジュールは、クライアントからマルチキャストされたファイル操作要求を共通に受け取り、当該要求に応じて自サーバの前記マッピングテーブルを参照することで、自サーバが当該要求を処理可能な最適なサーバであるか否かを判断し、最適なサーバであると判断した場合だけ、要求されたファイル操作を対応するサーバの前記ローカルファイルシステムにより行わせるように構成されていることを特徴とする仮想分散ファイルサーバシステム。

【請求項 2】 マルチキャスト可能なネットワーク上に分散した複数のファイルサーバを備えた仮想分散ファイルサーバシステムであって、

前記各ファイルサーバに分散して実装され、全ファイルサーバのファイルを統合的に管理する、実際のストレージ構成には非依存の仮想分散ファイルシステムと、

前記各ファイルサーバにそれぞれ独立して実装され、各サーバに固有のストレージ構成を管理するローカルファイルシステムと、

前記各ファイルサーバにそれぞれ設けられ、前記仮想分散ファイルシステムで統合的に管理される各ファイルについて、当該仮想分散ファイルサーバシステムとそのファイルを実際に管理する前記ローカルファイルシステムとの間のマッピングの情報を保持するマッピングテーブルと、

前記各ファイルサーバにそれぞれ設けられ、全ての前記ファイルサーバについて、そのサーバのストレージ装置の空き容量を示す情報、及びそのサーバの負荷状況を示す情報の少なくとも一方を含むサーバ情報を保持するサーバ情報保持手段とを具備し、

前記仮想分散ファイルシステムは、前記各ファイルサーバにそれぞれ設けられた管理モジュールから構成され、前記各管理モジュールは、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合、当

ストされたファイル操作要求を共通に受け取り、当該要求に応じて自サーバの前記マッピングテーブルまたは前記サーバ情報保持手段を参照することで、自サーバが当該要求を処理可能な最適なサーバであるか否かを判断し、最適なサーバであると判断した場合だけ、要求されたファイル操作を対応するサーバの前記ローカルファイルシステムにより行わせるように構成されていることを特徴とする仮想分散ファイルサーバシステム。

【請求項 3】 前記管理モジュールは、前記ファイル操作要求がファイル読み出し要求またはファイル書き込み要求の場合には、自サーバの前記マッピングテーブルを参照し、該当するファイルが自サーバの前記ローカルファイルシステムの管理下にあるか否かにより、自サーバが前記要求を処理可能な最適なサーバであるか否かを判断することを特徴とする請求項 1 または請求項 2 記載の仮想分散ファイルサーバシステム。

【請求項 4】 前記管理モジュールは、前記ファイル操作要求がファイルの新規作成要求の場合には、自サーバの前記サーバ情報保持手段を参照し、全ての前記サーバの各々について、そのサーバのストレージ装置の空き容量、またはそのサーバの負荷状況を比較することで、自サーバが前記要求を処理可能な最適なサーバであるか否かを判断することを特徴とする請求項 2 記載の仮想分散ファイルサーバシステム。

【請求項 5】 前記管理モジュールは、全ての前記ファイルサーバの前記マッピングテーブルの内容を一致化するために、自サーバの前記マッピングテーブルの情報と他のサーバの前記マッピングテーブルの情報とをサーバ間通信により交換することを特徴とする請求項 1 記載の仮想分散ファイルサーバシステム。

【請求項 6】 前記管理モジュールは、全ての前記ファイルサーバの前記マッピングテーブルの内容を一致化するために、自サーバの前記マッピングテーブルの情報と他のサーバの前記マッピングテーブルの情報とをサーバ間通信により交換する一方、全ての前記ファイルサーバの前記サーバ情報保持手段の内容を一致化するために、自サーバの前記サーバ情報保持手段の情報と他のサーバの前記サーバ情報保持手段の情報とをサーバ間通信により交換することを特徴とする請求項 2 記載の仮想分散ファイルサーバシステム。

【請求項 7】 前記各ファイルサーバにそれぞれ設けられ、そのサーバの管理下にある各ファイル別の負荷状況を示す情報を保持するファイル別負荷状況情報保持手段を更に具備し、

前記管理モジュールは、自サーバの前記ファイル別負荷状況情報保持手段に保持されている情報から第 1 の閾値を超えた負荷のファイルを検出して、他の任意のファイルサーバに対してサーバ間通信により当該ファイルのレプリケーションを行い、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合、当

該要求に対する処理をレプリケーション側に任せるようにしたことを特徴とする請求項 1 または請求項 2 記載の仮想分散ファイルサーバシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、コンピュータ・ネットワークシステムにおけるファイルサーバシステムに係り、特にネットワーク上に接続された複数のファイルサーバを協調動作させて、クライアントからは単一のサーバとして機能させる仮想分散ファイルサーバシステムに関する。

【0002】

【従来の技術】今日のコンピュータ・ネットワークシステムにおいては、ネットワークに接続された異なるコンピュータ間でファイルを共有することが一般的に行われている。こうした環境下では、特定のコンピュータに大規模なストレージを接続して、ファイルサーバとして運用したり、最近ではNAS (Network Attached Storage) と呼ばれる、ファイルサーバ専用機を接続する等のシステム構成をとることが多い。

【0003】ファイルサーバを使用する環境（ファイルサーバシステム）では、サーバのストレージ容量が不足した場合には、サーバ側に物理的・性能的に拡張性があれば、新たにディスク装置等（のストレージ装置）を増設することで対処できる。このときクライアントからは、新たなボリュームをマウントして使用するという形態になる。また、サーバの拡張性が限界に達していれば、サーバ自体を増設することになる。このときクライアントからは、増設したサーバを意識した上で新たなボリュームをマウントして使用するという形態になる。

【0004】

【発明が解決しようとする課題】上記したコンピュータ・ネットワークシステムにおいてファイルサーバを利用してファイル共有を行う場合、クライアント側からは、ファイルサーバ側のボリューム構成がそのまま見えてしまうのが一般的である。例えばサーバ側でディスク装置を増設した場合には、クライアント側は新たなボリュームを認識した上で、マウントしなければならない。或いはサーバ自体を増設した場合には、増設したサーバの運用ポリシーを決定、もしくはシステム設定・管理等の煩雑な作業が発生する上、クライアント側でも、新たなサーバを認識した上で、新たなボリュームをマウントしなければならない。

【0005】このように従来のファイルサーバを用いたファイル共有システム（ファイルサーバシステム）では、ディスク装置（ストレージ装置）の増設、或いはサーバの増設が必要な場合、サーバ側、クライアント側のいずれにも、新たな設定・管理のために多大なコストが発生するという問題があった。更に、ストレージの利用形態によっては、特定のファイルシステムをそのまま容量

だけ拡張したい場合もあり、単にストレージ装置やサーバを増設するだけでは解決しないケースもあった。

【0006】本発明は上記事情を考慮してなされたものでその目的は、ネットワーク上に分散した複数のファイルサーバを、クライアントからは単一のサーバとして扱うことができ、サーバ台数やストレージ装置の接続状態をクライアントに意識させない仮想分散ファイルサーバシステムを提供することにある。

【0007】

【課題を解決するための手段】本発明は、マルチキャスト可能なネットワークに接続された複数のファイルサーバに分散して実装され、全ファイルサーバのファイルを統合的に管理する、実際のストレージ構成には非依存の仮想分散ファイルシステムと、各ファイルサーバにそれぞれ独立して実装され、各サーバに固有のストレージ構成を管理するローカルファイルシステムと、前記各ファイルサーバにそれぞれ設けられ、上記各ファイルについて、仮想分散ファイルサーバシステムとそのファイルを実際に管理するローカルファイルシステムとの間のマッピングの情報（例えば、仮想分散ファイルサーバシステムで管理され、クライアントから見える仮想的なパスと、ローカルファイルシステムで管理され、クライアントから見えない物理的な所在とを対応付けた情報）を保持するマッピングテーブルとを備えると共に、上記仮想分散ファイルシステムを、各ファイルサーバにそれぞれ設けられた管理モジュールであって、クライアントからマルチキャストされたファイル操作要求を共通に受け取り、当該要求に応じて自サーバのマッピングテーブルを参照することで、自サーバが当該要求を処理可能な最適なサーバであるか否かを判断し、最適なサーバであると判断した場合だけ、要求されたファイル操作を対応するサーバのローカルファイルシステムにより行わせる管理モジュールにより構成することを特徴とする。

【0008】ここで、各ファイルサーバ上に、全ファイルサーバについて、そのサーバのストレージ装置の空き容量を示す情報、及びそのサーバの負荷状況を示す情報の少なくとも一方を含むサーバ情報を保持するサーバ情報保持手段を更に設け、上記各管理モジュールでは、クライアントからマルチキャストされたファイル操作要求を受け取った場合に、当該要求に応じて自サーバのマッピングテーブルまたはサーバ情報保持手段を参照することで、自サーバが当該要求を処理可能な最適なサーバであるか否かを判断する構成としてもよい。

【0009】このような構成においては、クライアントから特定のファイルサーバを意識することなくマルチキャストされたファイル操作要求は、仮想分散ファイルサーバシステムを構成する各ファイルサーバ上の管理モジュールで共通に受け取られ、その要求に応じて対応するサーバ（自サーバ）のマッピングテーブルまたはサーバ情報保持手段が参照される。そして、この参照の結果、

自サーバが上記要求を処理可能な最適なサーバであるか否かが判断され、最適なサーバであると判断した唯一のサーバ（上の管理モジュール）だけが、要求されたファイル操作を自サーバのローカルファイルシステムにより行わせる。

【0010】このように、要求元のクライアントからは、ネットワーク上に分散した複数のファイルサーバを単一のサーバとして扱うことができ、サーバ台数やストレージ装置の接続状態を意識する必要がない。

【0011】ここで、上記管理モジュールで、自サーバが最適なサーバであるか否かを判断するためのアルゴリズムとして、以下の第1乃至第4のアルゴリズム（判断手法）のいずれかを適用するとよい。

【0012】第1のアルゴリズムは、ファイル操作要求がファイル読み出し要求またはファイル書き込み要求の場合に適用されるもので、自サーバのマッピングテーブルの情報に基づいて、該当するファイルが自サーバのローカルファイルシステムの管理下にあるか否かにより判断する手法である。

【0013】第2のアルゴリズムは、ファイル操作要求がファイルの新規作成要求の場合に適用されるもので、自サーバのサーバ情報保持手段の情報に基づいて、全てのサーバの各々について、そのサーバのストレージ装置の空き容量（空き記憶容量）、またはそのサーバの負荷状況を比較することで判断する（例えば、自サーバの空き容量が最も大きい場合、或いは自サーバの負荷が最も低い場合に上記最適サーバと判断する）手法である。

【0014】第3のアルゴリズムも、ファイル操作要求がファイルの新規作成要求の場合に適用されるもので、自サーバのマッピングテーブルの情報に基づいて、全てのサーバの各々について対応するストレージ装置上に確保可能な連続領域を求め、その連続領域のサイズを比較することで判断する（例えば、自サーバのストレージ装置上に確保可能な連続領域のサイズが最も大きい場合に上記最適サーバと判断する）手法である。

【0015】第4のアルゴリズムも、ファイル操作要求がファイルの新規作成要求の場合に適用されるもので、全てのサーバの各々について、そのサーバのストレージ装置の空き容量、そのサーバの負荷、及び当該ストレージ装置上に確保可能な連続領域の少なくとも2つを求め、その求めた少なくとも2つの情報を複合条件として比較することで判断する手法である。

【0016】以上の第1乃至第4のアルゴリズムのいずれか1つを適用することで、クライアントから特定のファイルサーバを意識することなくマルチキャストされたファイル操作要求を各サーバが共通に受け取っても、その要求されたファイル操作を行うのに最適なサーバであるか否かを、その都度相互に通信を行うことなく、そのサーバ自身で自律的に判断することができる。

【0017】ここで、上記各管理モジュールに、全ての

ファイルサーバのマッピングテーブルの内容を一致化するために、自サーバのマッピングテーブルの情報と他サーバのマッピングテーブルの情報とをサーバ間通信により交換する機能（通信モジュール）を持たせるとよい。また、マッピングテーブルに加えてサーバ情報保持手段を各サーバ上に備えた構成では、各管理モジュール（内の通信モジュール）に、全てのファイルサーバのサーバ情報保持手段の内容を一致化するために、自サーバのサーバ情報保持手段の情報と他のサーバのサーバ情報保持手段の情報とをサーバ間通信により交換する機能を更に持たせるとよい。

【0018】また、マッピングテーブルの一致化のためには、自サーバのローカルファイルシステムで実際に管理されるファイル構成が変更された場合に、その変更された情報（マッピング情報）をサーバ間通信により他の全サーバに送信するのが効率的である。同様に、サーバ情報保持手段の内容の一致化のためには、自サーバのサーバ情報を定期的に更新し、その都度、その更新されたサーバ情報をサーバ間通信により他の全サーバに送信するのが効率的である。

【0019】また本発明は、上記仮想分散ファイルサーバシステムにサーバが動的に増設された場合に、そのサーバの管理モジュールで以下の第1乃至第4の処理を行うようにしたことをも特徴とする。まず、第1の処理では、サーバ間通信により他の全てのサーバに対してマッピングテーブル及びサーバ情報保持手段の更新を禁止するロック設定を行い、次の第2の処理では、サーバ間通信により他の任意のサーバからマッピングテーブルサーバ情報保持手段の内容を自サーバにコピーし、次の第3の処理では、自サーバのサーバ情報保持手段に自サーバのサーバ情報を追加し、次の第4の処理では、サーバ間通信により自サーバのサーバ情報を他の全てのサーバのサーバ情報保持手段に反映させて全サーバのサーバ情報保持手段の一致化を図り、しかる後に上記ロック設定を解除する。

【0020】このようなサーバ増設時の一連の動作により、動的にサーバ台数を拡張できる。しかもクライアントは、サーバ台数の拡張を意識することなく、増設されたサーバを利用することができる。

【0021】また本発明は、各ファイルサーバに、そのサーバの管理下にある各ファイル別の負荷状況を示す情報を保持するファイル別負荷状況情報保持手段を付加し、各サーバの管理モジュールにおいて、自サーバのファイル別負荷状況情報保持手段に保持されている情報から第1の閾値を超えた負荷のファイルを検出して、他の任意のファイルサーバに対してサーバ間通信により当該ファイルのレプリケーションを行い、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合、当該要求に対する処理をレプリケーション側に任せるようにしたことをも特徴とする。

【0022】このような構成においては、自律的な負荷分散が可能となる。ここで、上記検出したファイルがレプリケーションされたファイルである場合にも、他の任意のファイルサーバに対してサーバ間通信により当該ファイルのレプリケーションを行い、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合に、当該要求に対する処理を新たなレプリケーション側に任せることにより、自律的な負荷分散がより広範囲に効果的に行える。また、他サーバへのレプリケーションの対象となったファイルの負荷が前記第1の閾値より低い第2の閾値以下となった場合に、そのレプリケーションを行ったサーバ上の管理モジュールから当該他サーバに対してサーバ間通信により対応するファイルの消去を要求し、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合、当該要求に対する処理を自身が行うことにより、動的な負荷分散が可能となる。

【0023】さて、上記一致化のためのサーバ間通信（マッピング情報またはサーバ情報の通信）、更には上記レプリケーションのためのサーバ間通信には、上記ネットワークを用いることが可能である。しかし、各ファイルサーバを相互接続する専用の通信路（プライベート通信路）を上記ネットワークから独立に設け、当該通信路を用いてサーバ間通信を行う構成とするとよい。この場合、サーバ間通信のためにネットワークのスループットが悪化するのを防止できる。

【0024】また本発明は、各ファイルサーバ及び当該サーバのストレージ装置を相互接続するマルチホストが可能なインタフェースを更に備えると共に、上記各管理モジュールによる上記サーバ間通信を当該インタフェースを介して行うようにしたことをも特徴とする。

【0025】このような構成においては、上記各ストレージ装置を上記インタフェースによって各サーバ間で共有し、上記一致化のためのサーバ間通信、更には動的なファイルのレプリケーションのためのサーバ間通信が上記インタフェースを通して行われるため、自律的な負荷分散が効果的に実現される。

【0026】なお、本発明は方法に係る発明としても成立する。

【0027】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。

【0028】〔第1の実施形態〕図1は本発明の第1の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図である。

【0029】同図において、1は仮想分散ファイルサーバシステム、2は同ファイルサーバシステム1にファイルサービスを要求するクライアント（クライアントコンピュータ）である。仮想分散ファイルサーバシステム1

は、ネットワーク3上に分散配置された複数、例えば2台のファイルサーバ（サーバコンピュータ）10-1、10-2を用いて実現される。なお、図ではクライアント2は便宜的に1台だけが示されているが、複数存在するのが一般的である。

【0030】11は仮想分散ファイルサーバシステム1の中心をなす仮想分散ファイルシステムであり、各ファイルサーバ10-1、10-2に分散して実装されている。この仮想分散ファイルシステム11は、全ファイルサーバ10-1、10-2のファイルを統合的に管理し、ファイルサーバ10-1、10-2それぞれの実際のボリューム構成（ストレージ構成）には依存しない、仮想的なファイルシステムをクライアント2に対して提供するものである。

【0031】仮想分散ファイルシステム11は、各ファイルサーバ10-1、10-2に分散して実装された仮想分散ファイルモジュール110-1、110-2を有している。仮想分散ファイルモジュール110-1、110-2は、ファイルサーバ10-1、10-2上でクライアント2からの要求を分散して処理しつつ、クライアント2に対しては仮想的に1つのファイルシステムとして見せるための管理モジュールである。仮想分散ファイルモジュール110-1、110-2は、当該モジュール110-1、110-2の中心をなし、クライアント2からの要求を処理する仮想分散ファイルインタフェース111-1、111-2と、後述するローカルファイルシステム13-1、13-2とのインタフェース（ローカルファイルインタフェース）112-1、112-2と、当該モジュール110-1、110-2間での通信（後述するマッピングテーブル14-1、14-2の情報、及びサーバ情報の一致化のための通信に代表されるサーバ間通信）を行う通信モジュール113-1、113-2とを持つ。

【0032】ファイルサーバ10-1、10-2はネットワーク3を介してクライアント2と接続されている。ファイルサーバ10-1、10-2は、仮想分散ファイルシステム11の他に、それぞれ当該サーバ10-1、10-2に接続されたディスク装置などのストレージ装置12-1、12-2（実際のストレージ構成）を管理するローカルなファイルシステム（ローカルファイルシステム）13-1、13-2を実装している。

【0033】ファイルサーバ10-1、10-2には、仮想分散ファイルシステム11とローカルファイルシステム13-1、13-2とを対応付ける同一内容のマッピングテーブル14-1、14-2が設けられている。このテーブル14-i（i=1、2）のデータ構造を図2に示す。

【0034】テーブル14-iの各エントリは、仮想分散ファイルシステム1が管理するファイルのファイル名の登録フィールド（ファイル名フィールド）141と、当該ファイルの仮想分散ファイルシステム1上の論理的な所在を表す（クライアント2から見える）パス（仮想パ

ス)の登録フィールド(仮想パスフィールド)142、当該ファイルのストレージ上の物理的な所在位置を表す(クライアント2から見えない)所在位置情報の登録フィールド(所在位置情報フィールド)143、当該ファイルへのアクセス権(許可/禁止)を管理するためのパーミッション情報の登録フィールド(パーミッション情報フィールド)144、及びその他の各種属性の登録フィールド145を有している。

【0035】仮想分散ファイルシステム11(上の仮想分散ファイルモジュール110-i)は、このようなデータ構造のマッピングテーブル14-iを参照することにより、例えばあるファイルがファイルサーバ10-1、10-2のいずれにあるか等の所在情報を得ることができる。他、パーミッション等、必要に応じてファイルの属性を得ることができる。

【0036】ファイルサーバ10-1、10-2には更に、同一内容のサーバ情報保持部15-1、15-2が設けられている。サーバ情報保持部15-i(i=1,2)は、図3に示すように、仮想分散ファイルサーバシステム1を構成する全てのファイルサーバ10-1、10-2の(ストレージ装置12-1、12-2の)空き記憶容量を示す情報(リソース情報)、及び負荷状況を示す情報を含むサーバ情報を保持するのに用いられる。

【0037】次に図1の構成の動作を説明する。本実施形態では、クライアント2からは、各ファイルサーバ10-1、10-2のローカルファイルシステム13-1、13-2ではなくて、仮想分散ファイルシステム11がマウントされているように見えている。そこでクライアント2は、何らかのファイル操作要求が発生した場合、仮想分散ファイルシステム11が実装されている全ファイルサーバ10-1、10-2に対して同一の要求を発行する。この場合、例えばIP(Internet Protocol)マルチキャストを使用する等の手法によれば、クライアント2側はファイルサーバの台数を意識することなく要求の発行が可能である。

【0038】ファイルサーバ10-1、10-2は、クライアント2からの要求を受け取ると、当該要求を仮想分散ファイルシステム11内の自サーバに対応した仮想分散ファイルモジュール110-1、110-2に渡す。すると、モジュール110-1、110-2(内の仮想分散ファイルインタフェース111-1、111-2)は、その要求がファイルの読み出し要求もしくは書き込み(更新)要求であるか、または新規ファイルの作成要求もしくはディレクトリの作成要求であるか、その要求種別を判別する。

【0039】ここで、クライアント2からのファイル操作要求がファイルの読み出し要求もしくは書き込み要求であるものとする。この要求には、要求の対象となるファイルのファイル名と、当該ファイルの仮想分散ファイルシステム11上のパス(仮想パス)が付されている。

【0040】仮想分散ファイルモジュール110-1、110-2(内の仮想分散ファイルインタフェース111-1、111-2)は、クライアント2からのファイル操作要求がファイルの読み出し要求もしくは書き込み要求の場合、要求されたファイルのファイル名及び仮想パスにより自サーバ内のマッピングテーブル14-1、14-2を参照し、当該ファイル名及び仮想パスを持つテーブル14-1、14-2内エントリ中の所在位置情報フィールド143の登録情報から、操作要求のあったファイルが自サーバ内(自サーバに接続されたストレージ装置12-1、12-2)に保持されているか否かを調べる。

【0041】もし、要求されたファイルが自サーバ内に保持されている場合には、仮想分散ファイルモジュール110-1、110-2(内の仮想分散ファイルインタフェース111-1、111-2)は、ローカルファイルインタフェース112-1、112-2により自サーバ内のローカルファイルシステム13-1、13-2を介して実際のファイルにアクセスし、クライアント2に応答を返す。一方、操作要求のあったファイルが自サーバ内になかった場合には、他のサーバが応答するものと見なしして応答しない。

【0042】これに対し、クライアント2からの要求が新規ファイルの作成、或いはディレクトリの作成であった場合には、仮想分散ファイルモジュール110-1、110-2は、(マッピングテーブル14-1、14-2ではなくて)サーバ情報保持部15-1、15-2を参照する。そして、サーバ情報保持部15-1、15-2に保持されている全サーバのサーバ情報をもとに、所定のアルゴリズムに従い、いずれか1つのサーバ10-i(iは1または2)上の仮想分散ファイルモジュール110-iだけが、仮想分散ファイルインタフェース111-iによりクライアント2からの要求を受け付ける。具体的には、サーバ10-i上の仮想分散ファイルモジュール110-iは、全サーバのサーバ情報の示す空き記憶容量を比較し、自サーバ10-i(のストレージ装置12-i)の空き記憶容量が最も大きいと判定できる場合に、クライアント2からの要求を受け付けるものとする。この場合、必ずしもサーバ情報中に負荷状況の情報を持たせる必要はない。

【0043】なお、全サーバのサーバ情報の示す負荷を比較し、自サーバの負荷が最も低い場合にクライアント2からの要求を受け付けるようにしてもよい。この場合、必ずしもサーバ情報中に対応するサーバの空き記憶容量の情報を持たせる必要はない。

【0044】この他に、マッピングテーブル14-iの各ファイル毎のマッピング情報から(つまり各ストレージ装置12-1、12-2の領域の使用状況から)、各ストレージ装置12-1、12-2上に確保可能な連続領域を求め、必要なサイズ以上の連続領域が確保でき、且つそのサイズが最も大きいストレージ装置が自サーバのストレージ装置12-iの場合に、クライアント2からの要求を

受け付けるようにしてもよい。この場合、サーバ情報保持部15-1、15-2は必ずしも必要でない。

【0045】更に、空き記憶容量と負荷状況と確保できる連続領域のサイズの少なくとも2つを条件（複合条件）として評価値を求め、自サーバが要求を受け付ける最適なサーバであるか否かを判断ようにしてもよい。

【0046】さて、ファイルサーバ10-i上の仮想分散ファイルモジュール110-iでは、仮想分散ファイルインタフェース111-iによりクライアント2からの要求を受け付けると、要求された新規ファイルの作成、或いはディレクトリの作成を、ローカルファイルインタフェース112-iを介してローカルファイルシステム13-iにより行い、マッピングテーブル14-1、14-2に該当するエントリ情報を登録する。

【0047】新規ファイルの作成、或いはディレクトリの作成が完了した後は、ファイルサーバ10-i上の仮想分散ファイルモジュール110-iでは、自サーバ上のマッピングテーブル14-iに登録した新たなエントリ情報を通信モジュール113-iによりネットワーク3を介して他の全てのサーバ10-j（jは1または2、但しj≠i）上の仮想分散ファイルモジュール110-jに送る。仮想分散ファイルモジュール110-j（内の仮想分散ファイルインタフェース111-j）は、仮想分散ファイルモジュール110-iから送られたマッピングテーブル14-iのエントリ情報を通信モジュール113-jを介して受け取る。そしてモジュール110-j（内のインタフェース111-j）は、受け取った他サーバ10-iのマッピングテーブル14-iのエントリ情報を自サーバ内のマッピングテーブル14-jに登録する。このように、ファイルサーバ10-1、110-2上の仮想分散ファイルモジュール110-1、110-2が相互にマッピングテーブル14-1、14-2の新規登録されたエントリ情報（更には更新されたエントリ情報）を交換し合うことで、当該マッピングテーブル14-1、14-2の内容の一致化を図ることができる。

【0048】また、ファイルサーバ10-1、10-2上の仮想分散ファイルモジュール110-1、110-2は、自サーバ上のサーバ情報保持部15-1、15-2に保持されている各サーバのサーバ情報のうち、自サーバのサーバ情報（空き記憶容量、及び負荷状況）を定期的に更新すると共に、その更新したサーバ情報を（通信モジュール113-1、113-2により）ネットワーク3を介して他の全てのサーバ（上の仮想分散ファイルモジュール110-2、110-1）に定期的に送ることで、各ファイルサーバ10-1、10-2のサーバ情報保持部15-1、15-2の内容の一致化を図る。つまり仮想分散ファイルモジュール110-1、110-2は定期的にサーバ情報を交換し合うことで一致化を図る。

【0049】以上の動作によって、ファイルサーバ10-1、10-2を自律的に分散・協調動作させることがで

き、クライアント2には実際にはファイルサーバが2台（複数台）あることを意識させずに、仮想的なファイルサーバを提供することができる。

【0050】なお、図1のシステムの例ではサーバが2台である場合について説明したが、サーバが3台以上であつても同様の仕組みによって、仮想的なファイルサーバを提供することができる。

【0051】〔第2の実施形態〕図4は本発明の第2の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図であり、図1と同一部分には同一符号を付してある。

【0052】図4において、仮想分散ファイルサーバシステム4の中心をなす仮想分散ファイルシステム41は、n台のファイルサーバ10-1～10-nに分散して実装されている。この仮想分散ファイルシステム41は、図1中の仮想分散ファイルシステム11と同様に、全ファイルサーバ10-1～10-nのファイルを統合的に管理し、各ファイルサーバ10-1～10-nそれぞれの実際のボリューム構成には依存しない、仮想的なファイルシステムをクライアント2に対して提供している。仮想分散ファイルシステム41は、各ファイルサーバ10-1～10-n上でクライアント2からの要求を処理する仮想分散ファイルモジュール410-1～410-nを有している。モジュール410-1～410-nは、図1中のモジュール110-1、110-2と同様の構成である、仮想分散ファイルインタフェース411-1～411-nと、ローカルファイルインタフェース412-1～412-nと、通信モジュール413-1～413-nとを持つ。但し、本実施形態における通信モジュール413-1～413-nは、図1中の通信モジュール113-1、113-2と異なり、後述するプライベート通信路5を介して通信を行うように構成されている。

【0053】ファイルサーバ10-1～10-nはネットワーク3を介してクライアント2と接続されている。ファイルサーバ10-1～10-nは、仮想分散ファイルシステム11の他に、それぞれ当該サーバ10-1～10-nに接続されたストレージ装置12-1～12-2を管理するローカルなファイルシステム（ローカルファイルシステム）13-1～13-2を実装している。ファイルサーバ10-1～10-nには、マッピングテーブル14-1～14-nと、サーバ情報保持部15-1～15-2とが設けられている。

【0054】図4の構成の仮想分散ファイルサーバシステム4の特徴は、図1の構成の仮想分散ファイルサーバシステム1と異なつて、システムを構成するファイルサーバの台数がn台である点と、そのn台のファイルサーバ10-1～10-nがネットワーク3とは別のプライベート通信路5によつても相互接続されている点である。このプライベート通信路5は、例えばイーサネット、或いはファイバチャネル（Fibre Channel）等であるが、

物理層に関しては特定しない。またトポロジに関しても、図4の例ではバス型を想定しているが、ループやスイッチであってもよい。

【0055】図4の構成において、ファイルサーバ10-1~10-nが（仮想分散ファイルシステム41内の仮想分散ファイルモジュール410-1~410-nにより）分散・協調動作を行うためには、前記第1の実施形態での動作説明から類推されるように、マッピングテーブル14-1~14-n、及びサーバ情報保持部15-1~15-nの内容を、各サーバ10-1~10-n間で常に一致化しておく必要がある。しかし、サーバ10-1~10-n間の情報一致化を、前記第1の実施形態と同様にネットワーク3を介して行うのでは、仮想分散ファイルサーバシステム4を構成するファイルサーバの台数が増加した場合には、その情報一致化の（ためのサーバ間通信の）トラフィックが増加し、ネットワーク3上のスループットを悪化させることになる。

【0056】そこで本実施形態（第2の実施形態）では、図4の構成のように、各ファイルサーバ10-1~10-n間にサーバ間の情報交換専用のプライベート通信路5を設け、仮想分散ファイルシステム41内の仮想分散ファイルモジュール410-1~410-nで通信モジュール413-1~413-nにより行われるサーバ間通信に、即ちマッピングテーブル14-1~14-n、及びサーバ情報保持部15-1~15-nの内容を一致化するためのサーバ間通信に、この通信路5を使用するようにしている。

【0057】このように本実施形態では、マッピングテーブル14-1~14-n、及びサーバ情報保持部15-1~15-nの内容の一致化のためのサーバ間通信に、ネットワーク3でなくてプライベート通信路5を用いることにより、ネットワーク3の負荷の軽減を図ることができる。

【0058】〔第3の実施形態〕以上に述べた第1、第2の実施形態では、複数のファイルサーバを分散・協調動作させる仮想分散ファイルサーバシステムの構成例を示した。この第1、第2の実施形態で参照した図1、図4の構成は、特定のサーバ台数における静的な例である。しかし、サーバ台数については、変更可能な構成とすることが好ましい。

【0059】そこで、仮想分散ファイルサーバシステムを構成するサーバ台数を動的に拡張可能とした本発明の第3の実施形態について図面を参照して説明する。図5は本発明の第3の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図であり、図4と同一部分には同一符号を付してある。

【0060】まず、図4に示した仮想分散ファイルサーバシステム4、即ちn台のファイルサーバ10-1~10-nで構成される仮想分散ファイルサーバシステム4に、図5（a）に示すように、新たなファイルサーバ10-

(n+1)を追加するものとする。

【0061】この場合、追加されたファイルサーバ10-(n+1)にも分散されている仮想分散ファイルシステム41上の仮想分散ファイルモジュール410-(n+1)は、既に仮想分散ファイルサーバシステム4を構成しているファイルサーバ10-1~10-nに対し、図5（a）において符号A1で示すように、（例えば図示せぬプライベート通信路を介しての）サーバ間通信により、マッピングテーブル14-1~14-nのエントリ情報及びサーバ情報保持部15-1~15-nのサーバ情報（各サーバのリソース情報及び負荷状況を含む）の更新をロックする。

【0062】その上で、追加されたサーバ10-(n+1)上のモジュール410-(n+1)は、他のファイルサーバ10-1~10-nのうちのいずれかのサーバ、例えばファイルサーバ10-1から、図5（b）において符号A2で示すように、マッピングテーブル14-1及びサーバ情報保持部15-1の全情報を、サーバ間通信により自サーバ内のマッピングテーブル14-(n+1)及びサーバ情報保持部15-(n+1)にコピーする。

【0063】次に、追加されたファイルサーバ10-(n+1)上のモジュール410-(n+1)は、コピー後のサーバ情報保持部15-(n+1)に対し、図5（c）において符号A3で示すように、自サーバのリソース及び負荷状況を示すサーバ情報を追加する。

【0064】しかる後にファイルサーバ10-(n+1)上のモジュール410-(n+1)は、図5（d）において符号A4で示すように、サーバ間通信により他の全ファイルサーバ10-1~10-nに対してサーバ情報の一致化要求を発行し、その後にロックを解除する。

【0065】以上の一連の動作により、既に構築されている仮想分散ファイルサーバシステム4に対して、動的に新たなサーバ（ファイルサーバ10-(n+1)）を追加することができる。この場合、例えば現在の仮想分散ファイルサーバシステム4のボリューム構成に対し、新規リソースをどのように振り分けるか、といった情報を付加すれば、必要に応じてボリュームを選択的に拡張することも可能である。

【0066】〔第4の実施形態〕図6は本発明の第4の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図であり、図4と同一部分には同一符号を付してある。

【0067】図6において、6は図4中の仮想分散ファイルサーバシステム4に相当する仮想分散ファイルサーバシステムである。この仮想分散ファイルサーバシステム6の特徴は、当該システム6を構成するファイルサーバ10-1~10-n内に、自サーバ（のストレージ装置12-1~12-n）に保持されている各ファイルについての負荷状況の情報（ファイル別負荷状況情報）を保持するファイル別負荷状況情報保持部16-1~16-nを備えて

いる点にある。これに伴い、仮想分散ファイルサーバシステム 6 の中心をなす仮想分散ファイルシステム 61 の持つ機能も図 4 中の仮想分散ファイルシステム 41 とは一部異なる。但し、仮想分散ファイルシステム 61 内の各ファイルサーバ 10-1~10-n 毎の仮想分散ファイルモジュールには便宜的に図 4 と同一符号 (410-1~410-n) を用いている。なお、図 6 では、モジュール 410-1~410-n 内の構成要素 (仮想分散ファイルインタフェース、ローカルファイルインタフェース、通信モジュール)、及びファイルサーバ 10-1~10-n 上のマッピングテーブル、サーバ情報保持部は省略されている。

【0068】ファイル別負荷状況情報保持部 16-i (i=1~n) は、図 7 (a) に示すデータ構造を持ち、自サーバ内のファイル毎の負荷状況を示す情報と、そのファイルの属性を示す情報 (ファイル属性) と、レプリケーションフラグとを含むファイル別負荷状況情報を保持する。ファイル属性は、対応するファイルがオリジナルであるかレプリカ (複製) であることを示す。また、レプリケーションフラグは、対応するフラグがオリジナルの場合、そのレプリカを他サーバ側に生成済みであるか否か、つまりレプリケーション済みであるか否かを示す。

【0069】図 6 には、ファイルサーバ 10-n が持つストレージ装置 12-n 内に、ファイルサーバ 10-1 が持つストレージ装置 12-1 内の任意のファイル 611 のレプリカ 612 がレプリケーション B1 により保持されている様子が示されている。

【0070】次に、図 6 の構成の動作を説明する。仮想分散ファイルシステム 61 上の各仮想分散ファイルモジュール 410-1~410-n は、ファイル別負荷状況情報保持部 16-1~16-n を例えば定期的に参照する。そしてモジュール 410-1~410-n は、保持部 16-1~16-n に保持されているファイル別の負荷状況情報から、自サーバ (のストレージ装置 12-1~12-n) に保持されているファイルの中に、第 1 の閾値を超えた負荷のファイルが存在することを検出した場合、他サーバの 1 つに対して対応するファイルのレプリカを非同期に生成するレプリケーション動作を、例えばプライベート通信路 5 を介してのサーバ間通信により行う。ここでファイルの負荷状況は、当該ファイルへの要求の待ち行列 (キュー) にある要求数、或いは当該ファイルの待ち状態にある要求の示すサイズの総和であり、要求を受け付ける毎と要求を処理し終える毎に更新される。また、レプリケーションの対象サーバには、図示せぬサーバ情報保持部に保持されているサーバ情報に基づいて、例えば負荷が最も低いサーバを選択すればよい。

【0071】仮想分散ファイルモジュール 410-1~410-n はレプリケーション動作を行うと、自サーバのファイル別負荷状況情報保持部 16-1~16-n に保持されている対応するファイルの負荷状況情報中のレプリケー

ションフラグをレプリケーション済みの通知状態にセットする。またレプリケーション動作の対象となったサーバの仮想分散ファイルモジュールは、自サーバのファイル別負荷状況情報保持部内に対応するレプリカの負荷状況情報を追加する。

【0072】ここでは、図 6 に示すように、ファイルサーバ 10-1 の保持するファイル 611 のレプリケーション B1 がプライベート通信路 5 を介してファイルサーバ 10-n に対して行われて、そのレプリカ 612 が当該ファイルサーバ 10-n のストレージ装置 12-n に保持されたものとする。この場合、ファイルサーバ 10-1 のファイル別負荷状況情報保持部 16-1 に保持されているファイル 611 の負荷状況情報中のレプリケーションフラグがレプリケーション済みを示す状態にセットされる。また、ファイルサーバ 10-n のファイル別負荷状況情報保持部 16-n には、ファイル 611 のレプリカ 612 についての新たな負荷状況情報が追加される。この負荷状況情報中のファイル属性は、対応するファイルが (ファイル 611 の) レプリカ (612) であることを示す。

【0073】以後、クライアント 2 からファイル 611 の新たな読み出し要求があった場合、当該ファイル 611 を保持するファイルサーバ 10-1 (の仮想分散ファイルモジュール 410-1) は、当該ファイル 611 の負荷が第 2 の閾値 (但し、第 2 の閾値 < 第 1 の閾値) を超えているか否かを調べ、超えているならば、クライアント 2 からの要求に応答しない。この場合、クライアント 2 からの要求に対しては、レプリケーションを受けたファイルサーバ 10-n が応答する。ここでファイルサーバ 10-n は、ファイルサーバ 10-1 が応答するか否かを考慮する必要はなく、要求されたファイル 611 のレプリカ 612 を有する限り、クライアント 2 に応答すればよい。

【0074】このように、クライアント 2 からのファイル 611 に対する新たな読み出し要求を、そのレプリカ 612 を用いてファイルサーバ 10-n が処理することで、そのファイル 611 を保持するファイルサーバ 10-1 では、それ以前に受け付けた当該ファイル 611 に対する読み出し要求の処理が進み、当該ファイル 611 の負荷が上記第 2 の閾値以下となる。するとファイルサーバ 10-1 上の仮想分散ファイルモジュール 410-1 は、ファイルサーバ 10-n 上の仮想分散ファイルモジュール 410-n に対して、ファイル 611 のレプリカ 612 を消去するための要求を例えばプライベート通信路 5 を介したサーバ間通信により送る。

【0075】この要求を受けたファイルサーバ 10-n 上の仮想分散ファイルモジュール 410-n は、既に受け付け済みの要求に対してのみレプリカ 612 を用いて処理を行い、しかる後にレプリカ 612 と対応する負荷状況情報を消去する。一方、ファイルサーバ 10-1 上の仮想分散ファイルモジュール 410-1 は、クライアント 2 か

らのファイル 611 に対する新たな読み出し要求があれば、それに対して応答する。

【0076】ところで、ファイルサーバ 10-1 からファイルサーバ 10-n へのファイル 611 のレプリケーションにより、当該ファイル 611 に対する読み出し要求をファイルサーバ 10-n で受け付けるようになった結果、ファイルサーバ 10-1 におけるファイル 611 の負荷が第 2 の閾値以下となる前に、ファイルサーバ 10-n における当該ファイル 611 のレプリカ 612 の負荷が第 1 の閾値を超えることがあり得る。

【0077】そこで、このような場合、今度はファイルサーバ 10-n がレプリカ 612 を用いて次の世代のレプリカを他の 1 つのサーバに生成し、即ちレプリケーションのレプリケーションを行い、そのサーバでファイル 611 に対する読み出し要求を処理させればよい。そのためには、ファイル別負荷状況情報保持部 16-i (i = 1 ~ n) に保持されるファイル毎の負荷状況情報に、図 7 (a) に示したような負荷状況とレプリケーションフラグに加えて、図 7 (b) に示すように、ファイルの世代情報を持たせるとよい。

【0078】この場合、ある世代のレプリカの負荷が上記第 2 の閾値以下に下がった時点で、当該レプリカを持つサーバから、そのサーバによるレプリケーションの対象となったサーバの持つ次世代のレプリカを消去する等の制御を行うことができる。このとき、レプリカの消去が要求されたサーバが、別のサーバに対して更に次世代のレプリカを生成している場合、その更に次世代のレプリカを消去するとよい。この他に、少なくともレプリカに関連したファイルの負荷状況情報については、前記したサーバ情報と同様に、各サーバ間の一致化を図ることにより、同一ファイル (レプリカを含む) について、負荷が最も低いファイルを持つサーバが、当該ファイルに対する読み出し要求に応答するようにしてもよい。

【0079】最近では、ビデオ、オーディオ等のストリーミングデータや、或いは WWW (World Wide Web) のコンテンツ等、基本的には読み出しが主で、比較的サイズが大きく、ある程度のレスポンス (場合によっては帯域保証) が必要なデータが増加しつつある。しかもこうしたデータは、短期的に見て特定のデータ (ファイル) にアクセスが集中するケースが想定されるため、レスポンスを確保するのが困難な場合もある。以上に述べた図 6 の構成は、こうした状況を想定したもので、特定のファイルにアクセスが集中した場合に、自動的に当該ファイルのレプリケーションを行うことで、当該ファイルへのアクセスを分散させることができるようにしている。この構成は、単に負荷分散だけでなく、例えば重要性の高いファイルのバックアップに利用することも可能である。

【0080】〔第 5 の実施形態〕図 8 は本発明の第 5 の実施形態に係る仮想分散ファイルサーバシステムを適用

するコンピュータ・ネットワークシステムの構成を示すブロック図であり、図 4 と同一部分には同一符号を付してある。

【0081】図 8 において、8 は図 4 中の仮想分散ファイルサーバシステム 4 に相当する仮想分散ファイルサーバシステムである。この仮想分散ファイルサーバシステム 8 の特徴は、ファイルサーバ 10-1 ~ 10-n、及びストレージ装置 12-1 ~ 12-n が、例えば FC-AL (Fibre Channel Arbitrated Loop) 80 により相互接続され、(ホストとしての) 各ファイルサーバ 10-1 ~ 10-n から (ターゲットとしての) ストレージ装置 12-1 ~ 12-n の共有が可能な (つまりマルチホスト可能な) ネットワーク構成を適用している点にある。ここでは、図 4 の構成と異なって、プライベート通信路 5 を持たない点に注意されたい。

【0082】この図 8 の構成では、図 4 の構成において (仮想分散ファイルモジュール 410-1 ~ 410-n の通信モジュール 413-1 ~ 413-n により) プライベート通信路 5 を介して行われるサーバ間通信を、図 1 の構成と同様にネットワーク 3 を介して行えばよい (図は、この状態が示されている)。また、上記サーバ間通信を、ファイルサーバ 10-1 ~ 10-n のストレージ接続用のインタフェースを介して FC-AL 80 上で行うようにしてもよい。この場合、プライベート通信路 5 を用いたのと同様に、ネットワーク 3 の負荷を軽減できる。

【0083】図 8 の構成によれば、ストレージ装置 12-1 ~ 12-n が全てのファイルサーバ 10-1 ~ 10-n から直接に見えるので、各サーバ 10-1 ~ 10-n に図 6 中のファイル別負荷状況情報保持部 16-1 ~ 16-n を持たせることで、前記第 4 の実施形態で述べたようなレプリケーション動作や負荷分散を容易に行うことができる。なお、マルチホスト可能なネットワーク (インタフェース) は FC-AL 80 に限るものではなく、SCSI (Small Computer System Interface) バスであっても構わない。

【0084】

【発明の効果】以上詳述したように本発明によれば、ネットワーク上に分散した複数のファイルサーバを、クライアントからは単一のサーバとして扱うことができ、サーバ台数やストレージ装置の接続状態をクライアントに意識させることがない。

【0085】また本発明によれば、サーバを増設した場合、動的にボリュームを拡張することもできる。

【0086】更に本発明によれば、複数のサーバ間で自律的な負荷分散が実現できる。

【図面の簡単な説明】

【図 1】本発明の第 1 の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図。

【図 2】図 1 中のマッピングテーブルのデータ構造例を

示す図。

【図3】図1中のサーバ情報保持部のデータ構造例を示す図。

【図4】本発明の第2の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図。

【図5】本発明の第3の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図。

【図6】本発明の第4の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図。

【図7】図6中のファイル別負荷状況情報保持部のデータ構造例を示す図。

【図8】同実施形態の動作を説明するタイミングチャート。

【符号の説明】

- 1, 4, 6, 8…仮想分散ファイルサーバシステム
2…クライアント

3…ネットワーク

5…プライベート通信路

10-1～10-n…ファイルサーバ

11, 41, 61…仮想分散ファイルシステム

12-1～12-n…ストレージ装置

13-1～13-n…ローカルファイルシステム

14-1～14-n…マッピングテーブル

15-1～15-n…サーバ情報保持部

16-1～6-n…ファイル別負荷状況情報保持部

80…F C - A L (マルチホスト可能なインタフェース)

110-1～110-n, 410-1～410-n…仮想分散ファイルモジュール (管理モジュール)

111-1～111-n…仮想分散ファイルインタフェース

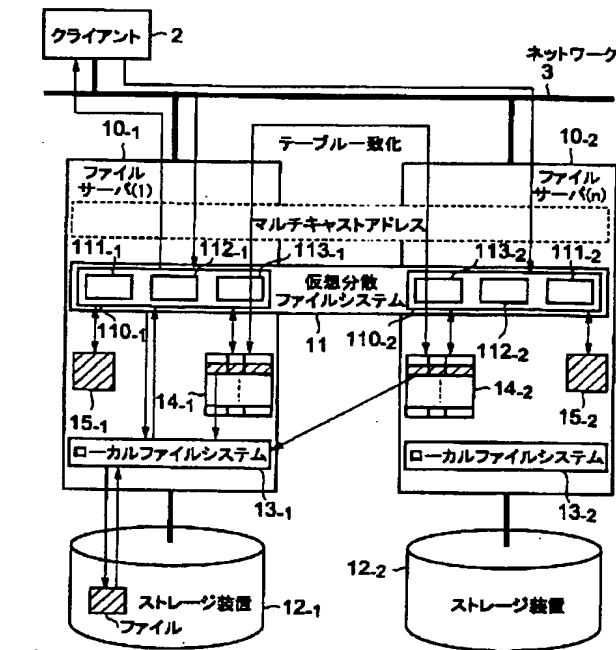
112-1～112-n…ローカルファイルインタフェース

113-1～113-n…通信モジュール

611…ファイル

612… (ファイル611の) レプリカ

【図1】



- 1
仮想分散ファイル
サーバシステム
- 14-1, 14-2…マッピングテーブル
15-1, 15-2…サーバ情報保持部
110-1, 110-2…仮想分散ファイルモジュール
111-1, 111-2…仮想分散ファイルインタフェース
112-1, 112-2…ローカルファイルインタフェース
113-1, 113-2…通信モジュール

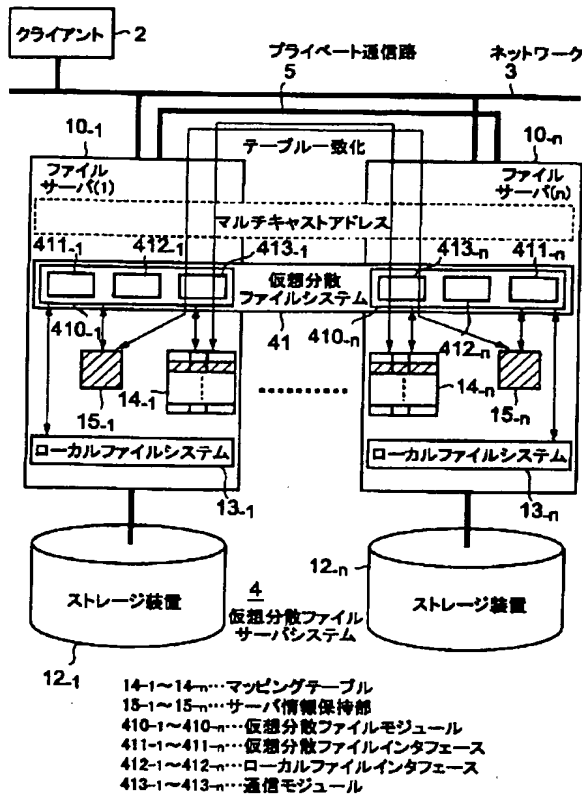
【図2】

141	142	143	144	145	145
ファイル名	仮想パス	所在位置情報	パーミッション情報	その他の属性	その他の属性
ファイル名	仮想パス	所在位置情報	パーミッション情報	その他の属性	その他の属性
ファイル名	仮想パス	所在位置情報	パーミッション情報	その他の属性	その他の属性

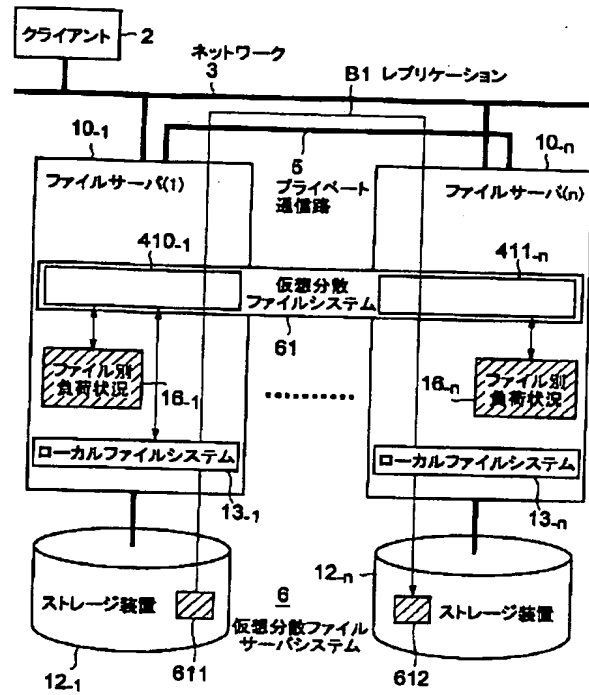
【図3】

サーバID	リソース情報	負荷状況
サーバ1		
サーバ2		

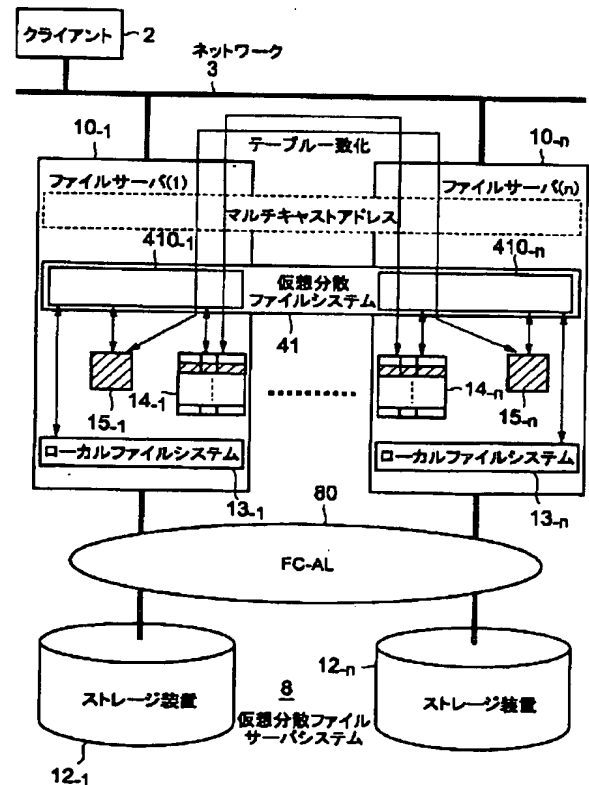
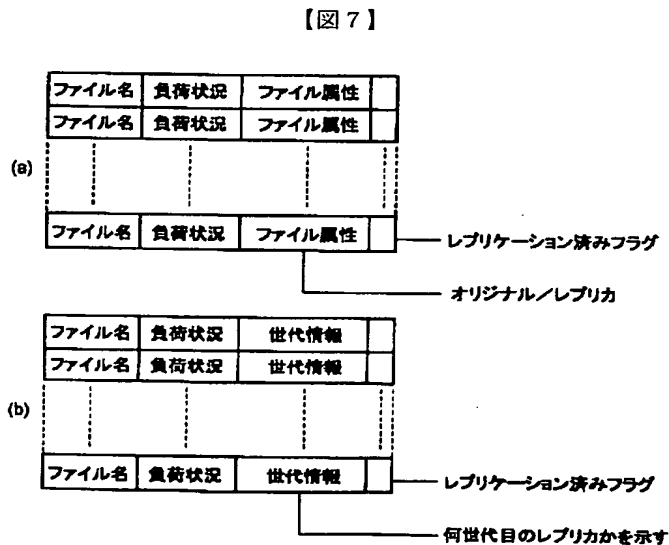
【図4】



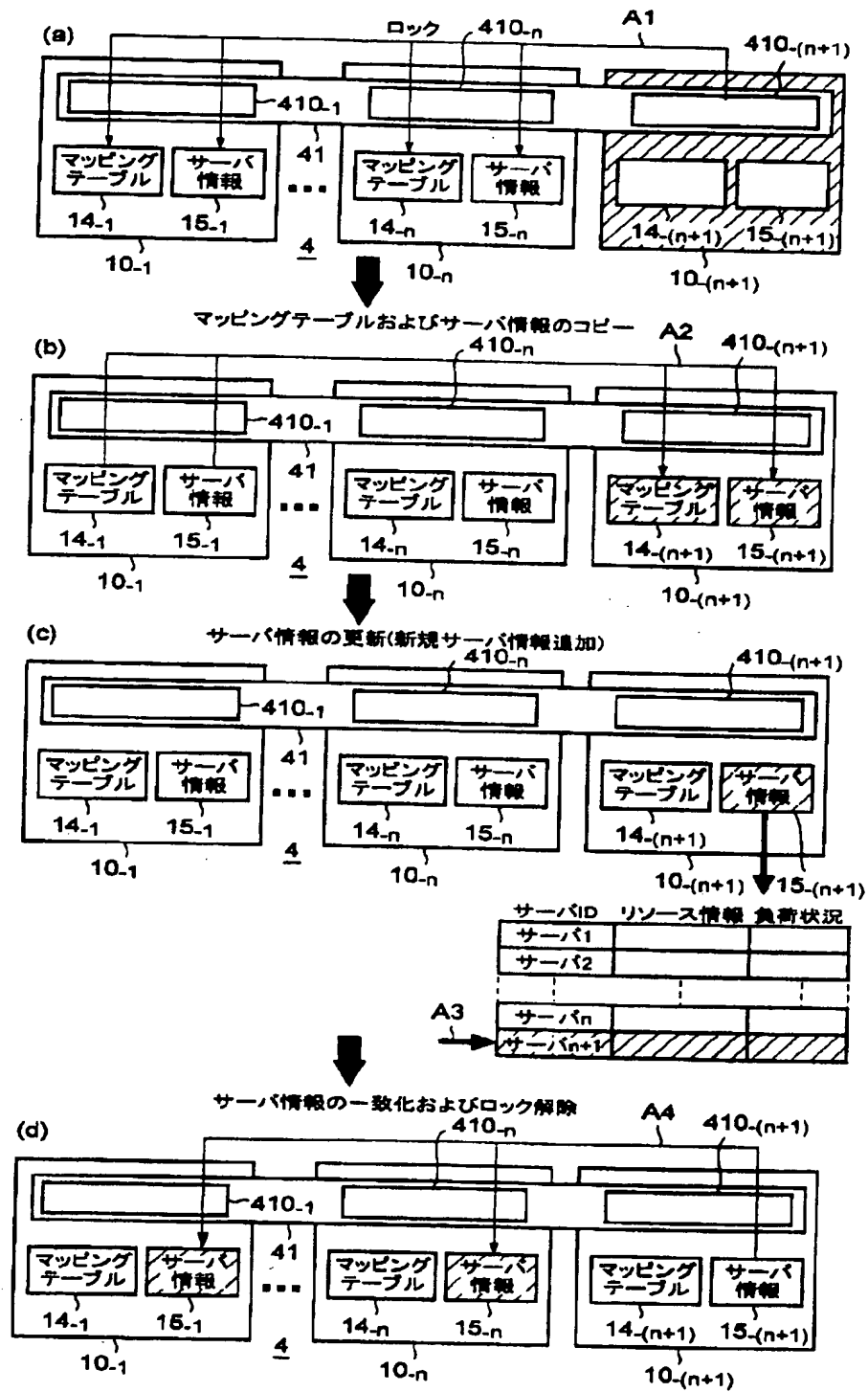
【図6】



【図8】



【図5】



フロントページの続き

F ターム(参考) 5B082 CA18 EA07 HA03 HA05 HA08
HA09
5B089 GA12 JA11 JB15 KA00 KC15
KC28 KE07